

Improved chromosome genome assembly, structural variation insight, and annotation of *Gossypium barbadense* L. 'Pima-S6'

Ricardo A. Chávez Montes¹, Mauricio Ulloa^{2,*}, Tammy Biniashvili³, Arik Zackay³, Nir Kfir³, Damar Lopez-Arredondo^{1,*}, and Luis Estrella-Herrera^{1,4,*}

¹ Institute of Genomics for Crop Abiotic Stress Tolerance, Plant and Soil Science Department, Texas Tech University, Lubbock, TX 79409, USA.

² USDA-ARS, PA, CSRL, Plant Stress and Germplasm Development Research, Lubbock, TX, 79415, USA.

³ NRGene Ness-Ziona, Israel

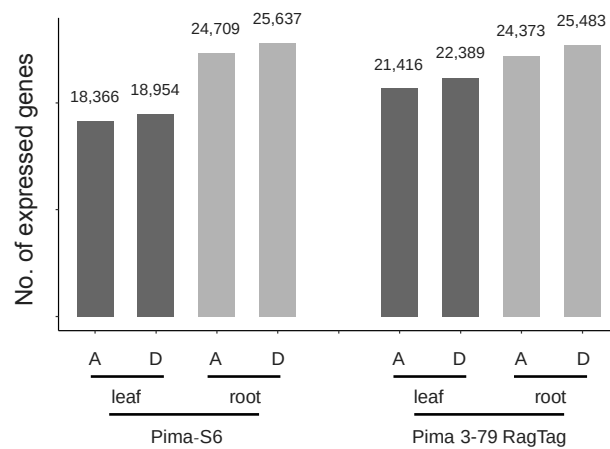
⁴ Unidad de Genómica Avanzada/Langebio, Centro de Investigación y de Estudios Avanzados, 36821, Irapuato, Mexico.

*Correspondence:

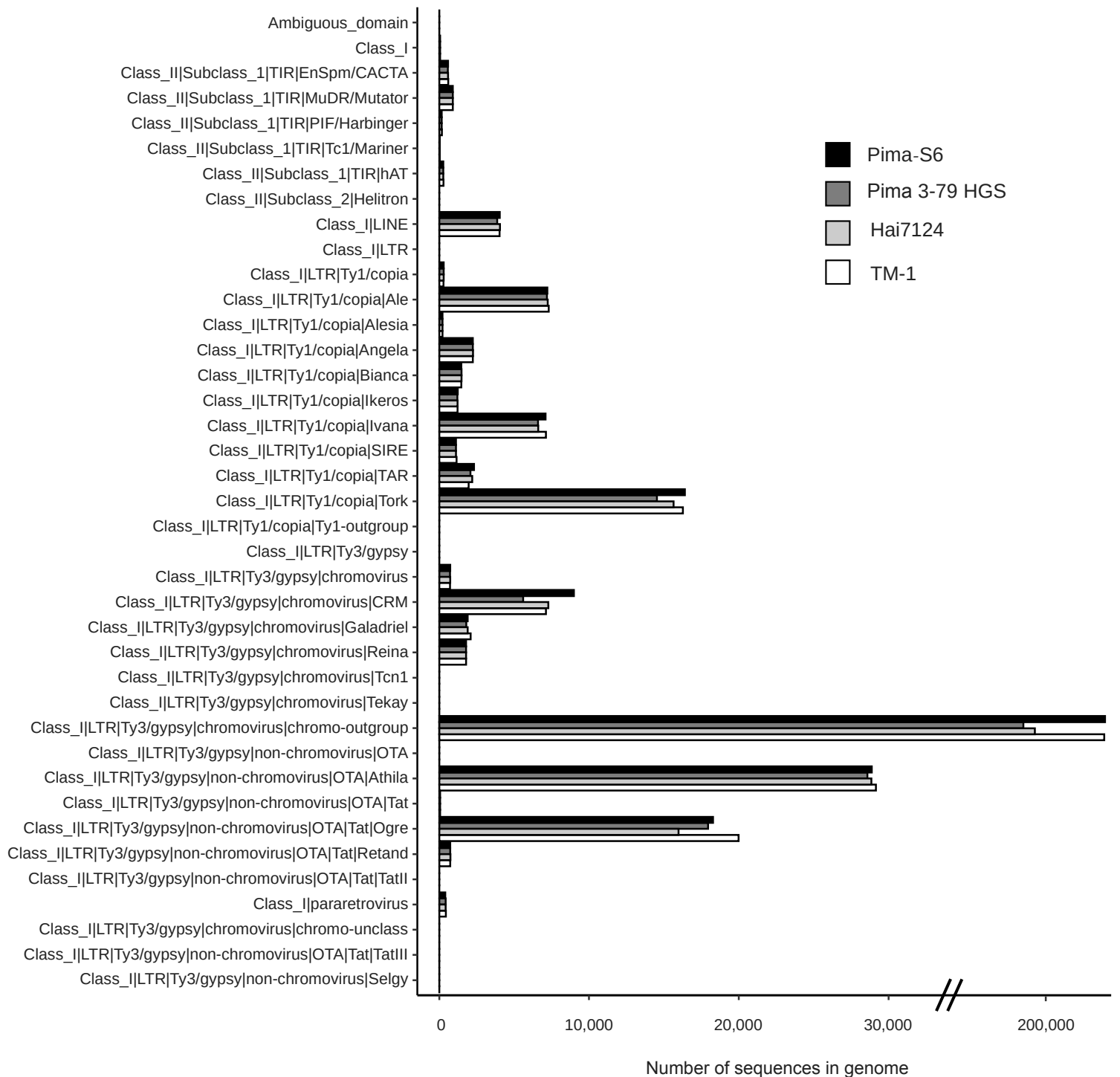
Luis Herrera-Estrella (Luis.Herrera-Estrella@ttu.edu)

Mauricio Ulloa (Mauricio.ulloa@usda.gov)

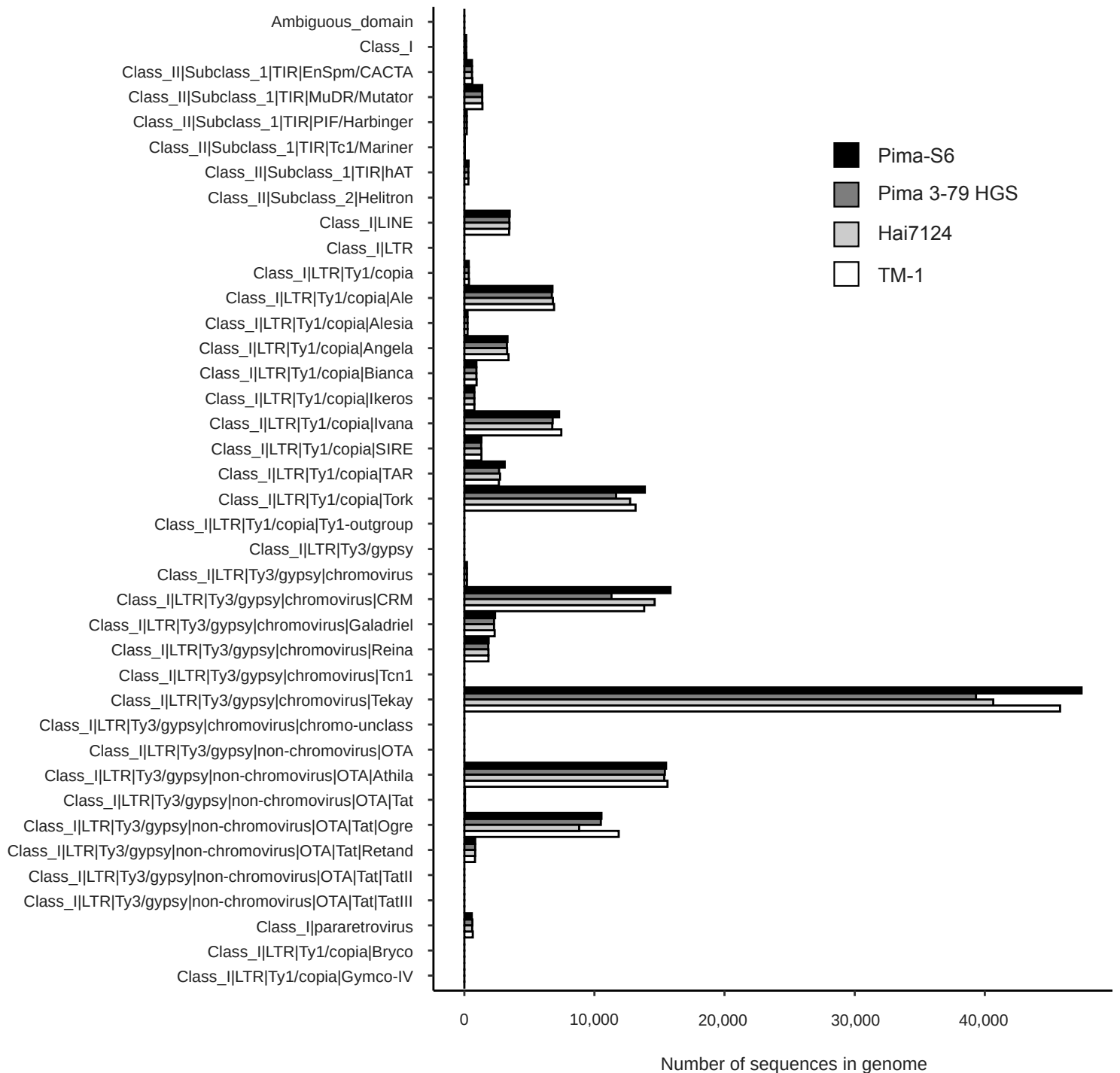
Damar Lopez-Arredondo (Damar.Lopez-Arredondo@ttu.edu)



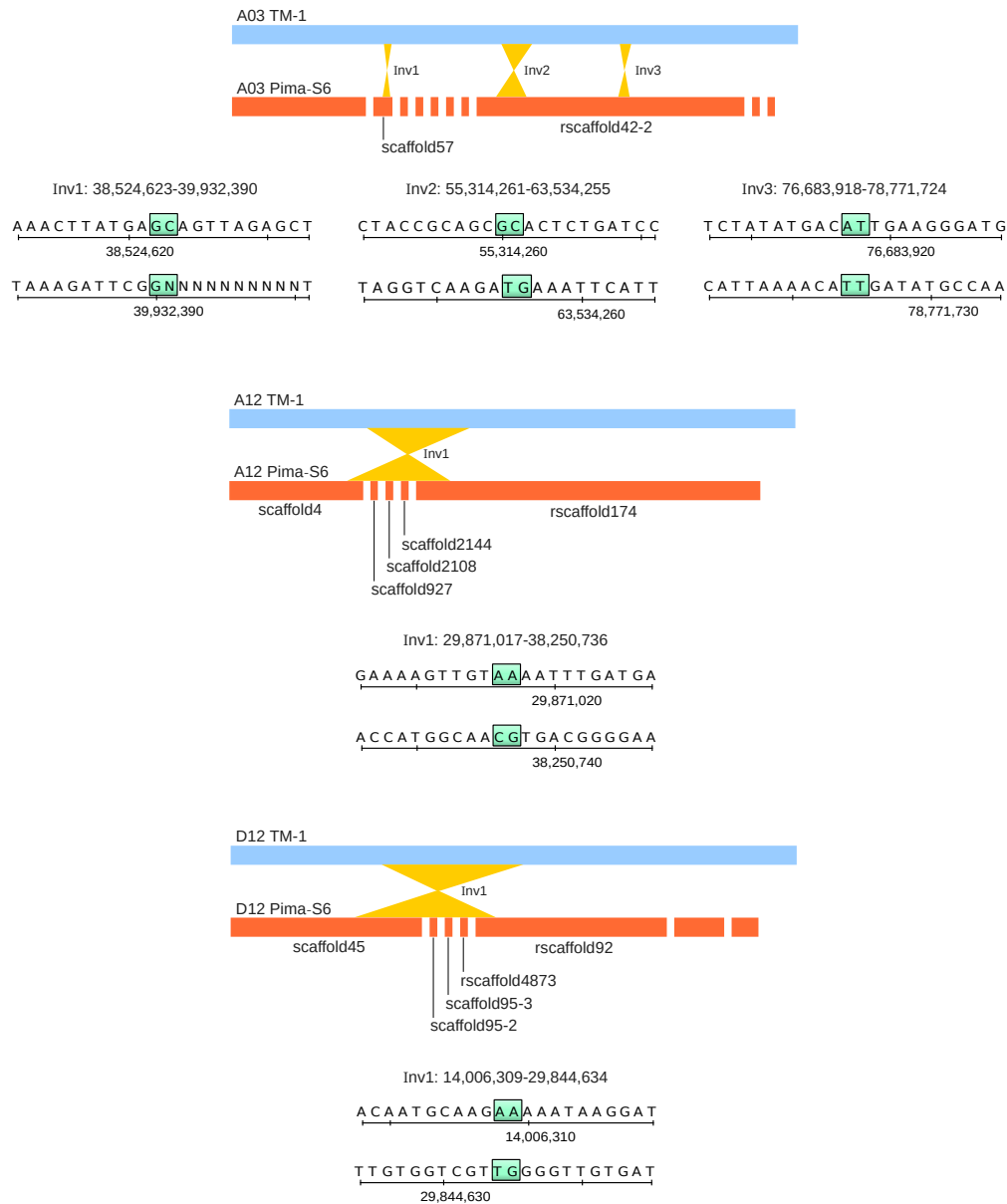
Supplementary fig. 1 Overview of leaf and root gene expression in Pima-S6 and Pima 3-79 RagTag. RNA from one leaf and one root sample was isolated and sequenced, and gene expression was quantified. Bars represent the number of expressed genes (TPM ≥ 1) in each species, organ and either their location on the A or D genome.



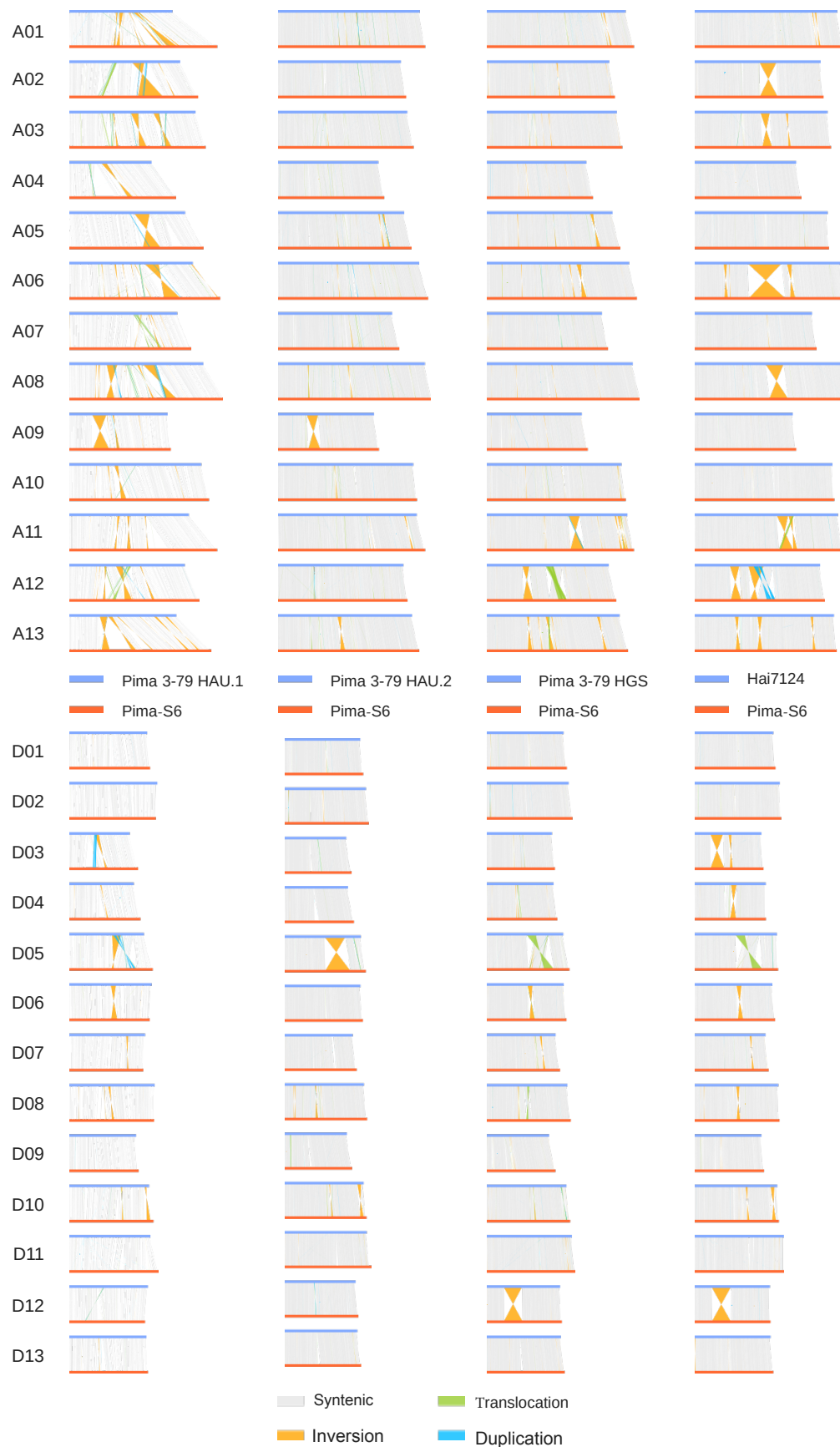
Supplementary fig. 2a Long Terminal Repeat (LTR) retrotransposon families distribution in four *Gossypium* assemblies in the A subgenome. The A13 subgenome chromosomes from Pima-S6, Pima 3-79 HGS, Hai7124 and TM-1 were analyzed using the Domain based ANnotation of Transposable Elements (DANTE) tool and the REXdb Viridiplantae v3.0 database. The output file was filtered on the RepeatExplorer Galaxy server at <https://repeatexplorer-elixir.cerit-sc.cz/> using default parameters, and the number of sequences per LTR retrotransposon family was plotted.



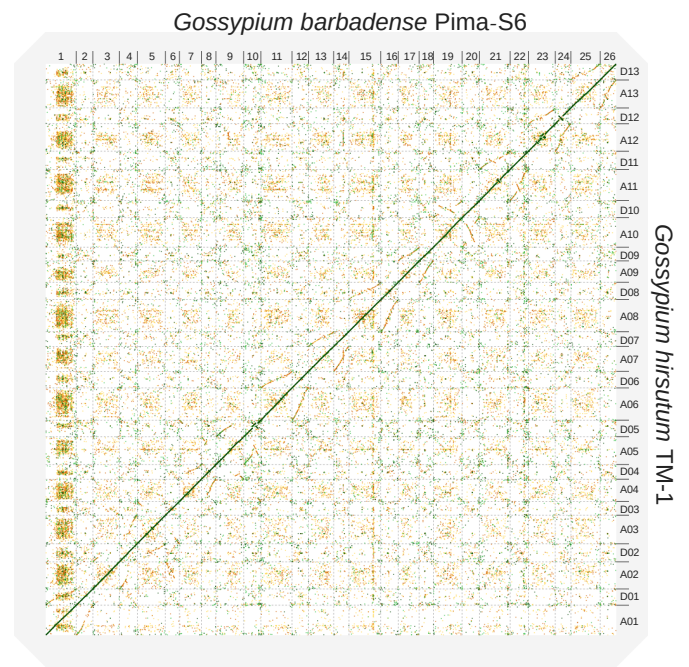
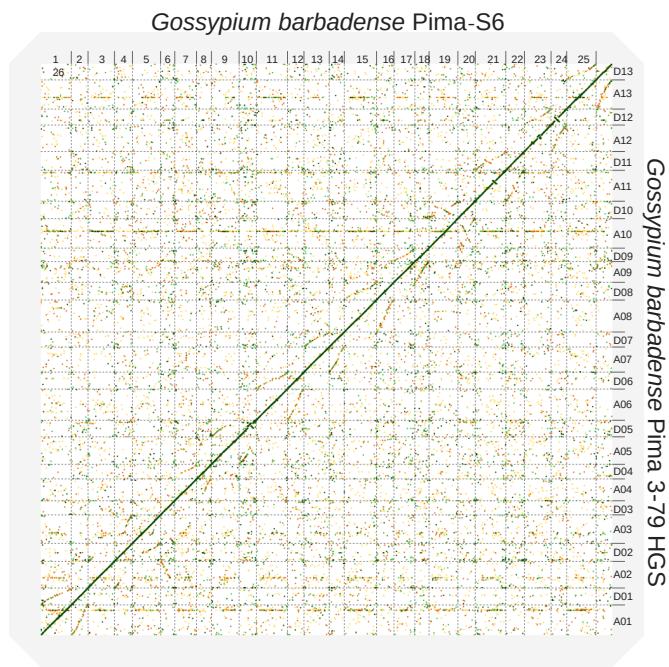
Supplementary fig. 2b Long Term Repeat (LTR) retrotransposon families distribution in four *Gossypium* assemblies in D subgenome. The D13 subgenome chromosomes from Pima-S6, Pima 3-79 HGS, Hai7124 and TM-1 were analyzed using the Domain based ANnotation of Transposable Elements (DANTE) tool and the REXdb Viridiplantae v3.0 database. The output file was filtered on the RepeatExplorer Galaxy server at <https://repeatexplorer-elixir.cerit-sc.cz/> using default parameters, and the number of sequences per LTR retrotransposon family was plotted.



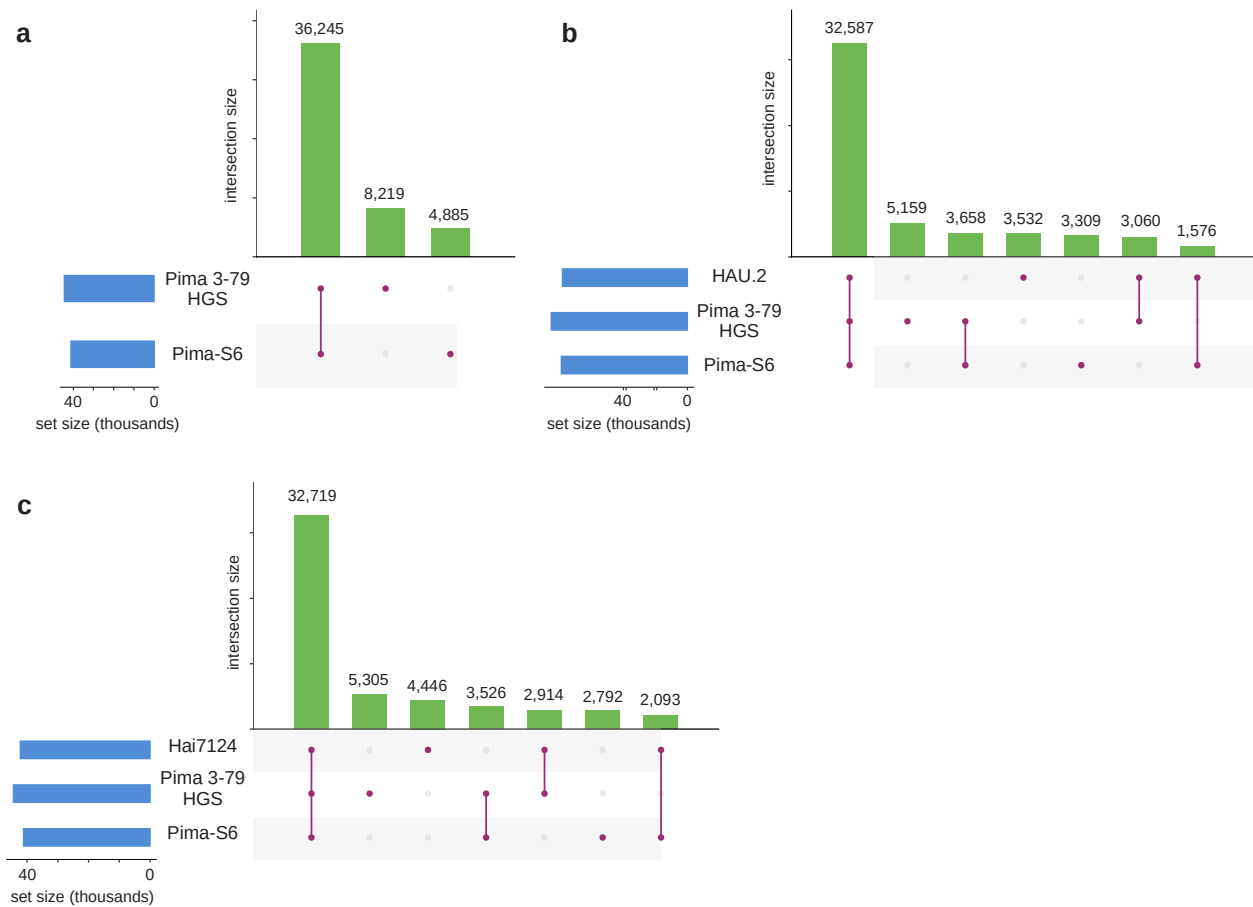
Supplementary fig. 3 TM-1 vs Pima-S6 synteny plots for chromosomes A03, A12 and D12. TM-1 chromosomes are represented by blue boxes and the scaffolds that form the corresponding Pima-S6 chromosomes are represented by red boxes. Relevant scaffold ids are indicated below each box. The prefix “r” in a scaffold name indicates that the scaffold was placed in reverse orientation. Gaps between scaffolds represent the 100 bp gaps introduced during chromosome reconstruction. Major inversions (> 1 Mbp) are represented by the orange hourglasses. For each inversion a close-up of the sequence at the inversion start and end is shown, with the exact inversion boundaries highlighted by green boxes. All inversion boundaries, except the end of Inversion 1 (Inv1) on chromosome A03, traverse unambiguously assembled regions. The 3' end of Inv1 on chromosome A03 is followed by a 10 bp N stretch and a 1,532 bp inversion-duplication of unambiguously assembled sequence that aligns to a chromosome A11 region on the TM-1 side.



Supplementary fig. 4 Synteny plots between Pima-S6 and Pima 3-79 HAU.1, Pima 3-79 HAU.2, Pima 3-79 HGS and Hai7124. In these synteny plots, the reference genome is represented by blue horizontal lines and the query genome by red horizontal lines. Vertical lines represent syntenic (gray), inverted (orange), translocated (green) and duplicated (blue) regions. Chromosome IDs are indicated on the left side of the plots. The 26 Pima-S6 chromosomes were aligned using minimap2 versus the 26 chromosomes of the Pima 3-79 HAU.1, Pima 3-79 HAU.2, Pima 3-79 HGS or Hai7124 genome assemblies, and synteny analysis was carried out using syri.



Supplementary fig. 5 Dot plots of Pima-S6 vs Pima 3-79 HGS and TM-1 genome assemblies alignments. The 26 chromosomes of the indicated genome assemblies were aligned using minimap2, and the PAF alignment file was plotted using D-Genies. Pima-S6 chromosome IDs were renamed to the AD nomenclature using the best matching chromosome in Pima 3-79 HGS and TM-1: chromosome 1 was renamed A01, chromosome 2 D01, chromosome 3 A02, chromosome 4 D02, and so on.



Supplementary fig. 6 Number of orthogroups shared between Pima-S6, Pima 3-79 HGS, Pima 3-79 HAU.2 and Hai7124. In these UpSet plots, set size indicates the number of orthogroups for each species, and intersection size indicates the number of orthogroups in each of the intersections shown below the corresponding bar. A total of 75,263 orthogroups were identified from the annotated proteins of *G. hirsutum* 'TM-1' (genome AD1), *G. barbadense* 'Pima-S6' (AD2; Pima-S6), *G. barbadense* 'Pima 3-79 HGS' (AD2; Pima 3-79 HGS), *G. barbadense* 'Pima 3-79' HAU.2 (AD2; HAU.2), *G. barbadense* Hai7124 (AD2; Hai7124), *G. tomentosum* (AD3), *G. mustelinum* (AD4), *G. darwinii* (AD5), *G. herbaceum* (A1), *G. arboreum* (A2) and *G. raimondii* (D5). **(a)** Number of orthogroups shared between Pima-S6 and Pima 3-79 HGS. The number of shared orthogroups is lower than the number of shared orthogroups between Pima-S6 and Pima 3-79 RagTag in Figure 5a. **(b)** Number of orthogroups shared between three Pima assemblies, Pima-S6, Pima 3-79 HGS and Pima-3-79 HAU.2. Pima 3-79 HGS and Pima 3-79 HAU.2 have an important number of proteins without an ortholog in any of the other two assemblies. **(c)** Number of orthogroups shared between Pima-S6, Pima 3-79 HGS and Hai7124. Pima 3-79 HGS has the highest number of orthogroups without orthologs in Pima-S6 or Hai7124.

Supplementary Table 1. Pima-S6 genomic sequencing libraries statistics.

Library type	Insert size	Sequencing	Number of libraries	Number of reads	Approximate depth
PCR-free paired-end	440-470 bp	2 x 250	1	370655026	160x
Mate-pair (Nextera™ MP Gel Plus)	2-4 kbp	2 x 150	1	222776057	58x
10X genomics™ Chromium™	NA	2 x 150	1	315583626	82x

Supplementary Table 2. Statistics for Pima-S6 and Pima 3-79 RNA-seq leaf and root RNA-seq samples.

Sample	Total reads	Aligned reads	Uniquely aligned reads	Aligned reads (%)	Uniquely aligned reads (%)
Pima-S6 leaf	4,105,123	2,858,840	2,407,125	69.6	58.6
Pima-S6 root	4,765,455	3,903,337	3,384,533	81.9	71.0
Pima 3-79 leaf	4,477,465	3,708,087	3,147,555	82.8	70.3
Pima 3-79 root	5,034,226	4,176,066	3,645,570	83.0	72.4

Supplementary Table 3. *G. barbadense* assemblies chromosome lengths. Chromosome lengths in bp of the Hai7124, Pima 3-79 HAU.1, Pima 3-79 HAU.2, Pima 3-79 HGS, and Pima 3-79 HGS-RagTag (our re-scaffolding of the Pima 3-79 HGS assembly) and Pima-S6 assemblies.

Chromosome ID	Hai7124	Pima 3-79 HAU.1	Pima 3-79 HAU.2	Pima 3-79 HGS	Pima 3-79 HGS-RagTag	Pima-S6
A01	115,639,457	83,454,191	115,637,255	113,238,469	114,571,553	120,119,894
A02	102,048,560	89,731,153	100,057,689	99,769,429	102,453,845	104,444,548
A03	107,689,529	102,275,589	105,315,579	105,981,974	107,161,519	110,526,773
A04	82,003,869	66,299,034	81,554,553	80,954,414	82,930,467	86,331,418
A05	107,778,512	93,810,923	102,776,486	102,458,744	105,538,069	108,864,071
A06	118,151,758	100,034,992	115,140,250	116,119,172	118,127,218	122,525,983
A07	95,011,077	87,365,090	92,880,876	93,754,744	95,177,537	98,767,759
A08	120,787,428	108,674,945	119,882,356	119,114,718	121,341,730	125,840,848
A09	79,226,513	79,449,896	77,927,517	77,140,527	78,896,073	82,042,076
A10	111,721,976	107,075,239	110,302,803	109,871,119	112,125,172	113,317,428
A11	116,189,812	96,770,036	113,101,708	114,694,469	117,157,560	120,065,767
A12	101,615,779	93,480,475	102,106,374	99,282,030	101,695,573	105,494,938
A13	112,976,367	86,578,892	109,233,746	108,235,369	110,287,702	115,221,231
A genome	1,370,840,637	1,195,000,455	1,345,917,192	1,340,615,178	1,367,464,018	1,413,562,734
D01	63,619,354	62,968,913	62,811,768	62,611,448	63,857,906	65,361,719
D02	68,884,445	71,317,061	67,659,327	66,877,699	67,695,607	70,102,368
D03	53,779,058	48,575,973	50,986,338	53,043,351	53,687,303	55,503,051
D04	57,348,712	51,938,909	52,479,919	54,342,187	54,661,041	57,477,808
D05	66,507,725	60,343,712	63,323,194	62,429,648	64,737,037	67,540,788
D06	62,791,478	66,700,274	62,964,862	62,820,931	63,351,879	65,076,424
D07	57,193,027	61,301,369	56,457,815	56,075,531	57,709,291	59,768,398
D08	67,735,652	69,128,419	65,987,480	65,831,471	66,566,626	68,651,644
D09	53,804,133	53,850,888	51,526,817	50,685,742	51,740,324	55,929,093
D10	66,844,369	64,637,854	65,706,892	65,066,651	66,412,153	68,149,462
D11	72,283,341	65,624,333	68,311,261	69,577,400	71,006,005	72,385,828
D12	60,787,228	63,542,810	58,873,029	59,787,526	59,929,623	61,230,550
D13	61,297,484	62,378,783	60,343,961	60,421,729	61,273,340	63,610,372
D genome	812,876,006	802,309,298	787,432,663	789,571,314	802,628,135	830,787,505
26 chromosomes	2,183,716,643	1,997,309,753	2,133,349,855	2,130,186,492	2,170,092,153	2,244,350,239